DECK 1 – INTRODUCTION TO DIGITIZATION AND ANALYTICS

Dr. Jack Hong
Adjunct Faculty, Lee Kong Chian School of Business, SMU
Co-founder, Research Room Pte. Ltd.

LESSON PLAN

LESSON PLAN

The materials in this pre-course are designed to equip you with:

- A high level understanding of the principles and application of data analytics
- Basic technical skills in implementing data analytics projects
  - Python and R

LESSON PLAN

- Deck 1 – Introduction
  - Demystifying digitization
  - Demystifying analytics
  - Analytics use-cases
  - Data science framework
  - Computing basics
- Deck 2 – Coding
  - Introduction to coding
  - Introduction to Python and R
- Deck 3 – Basic Statistical Models
  - Linear and Logistic Regression
- Deck 4 – Basic Machine Learning Models

DEMYSTIFYING DIGITIZATION

THE AGE OF DIGITIZATION

- The age of industrialization created technology and machines that scale our **physical abilities**
- The age of digitization created technology and machines that scale our **mental and decision making abilities**

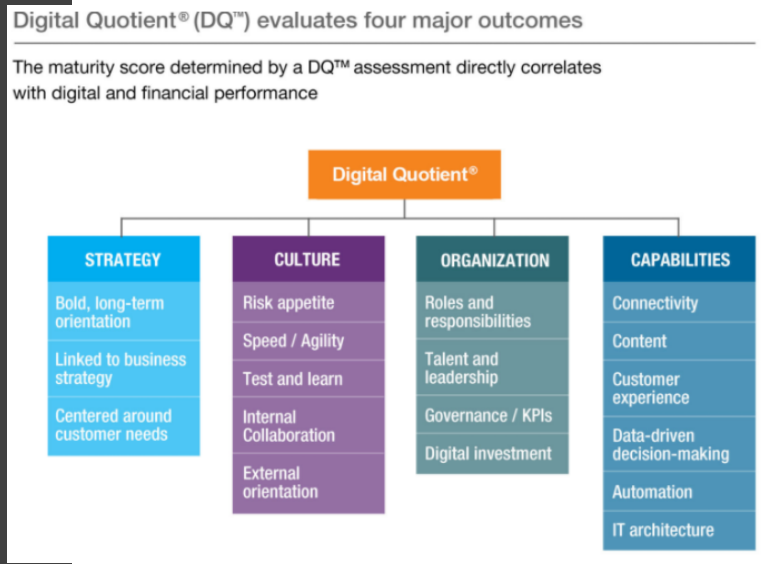
WHY DOES THE BUSINESS WORLD NEED TO CARE ABOUT DIGITIZATION?

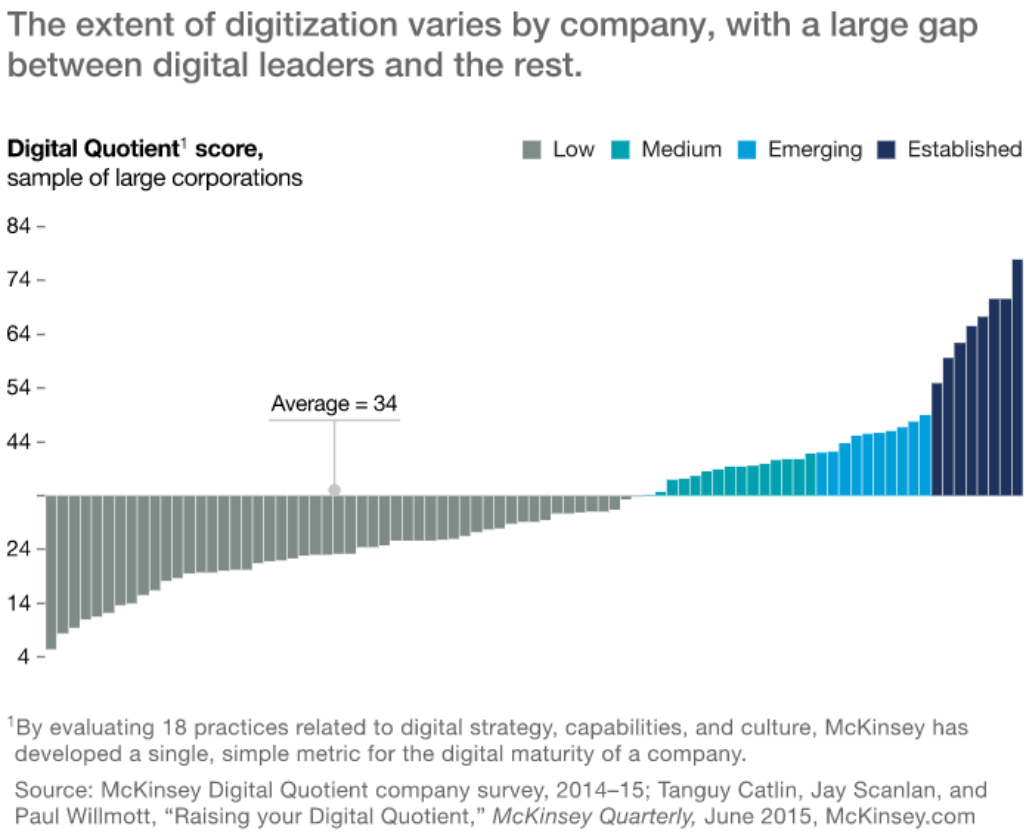
- Disrupt or be disrupted
  - Replicative (Provide more of existing goods and services: Cheaper, better, faster)
  - Innovative (Reshape industries and add value to the economy)
    - Google, Facebook, Wechat, Alibaba and Taobao
- Cater to changing consumer expectations or lose your customers
  - Instant gratification
  - News, anytime, anywhere
  - One click setup
  - 24/7 immediate response
  - Automated administrative tasks – e.g. approvals, transfers, form filling
  - Personalized treatment with seamless transition across multiple business or lifestyle touchpoints
  - Community wisdom and reputation building

WHY DO WE CARE ABOUT DIGITIZATION?

- Increase productivity
  - Reduce processes and increase accuracy (e.g. smart document identification, automated rules)
  - Identify anomalies accurately and instantaneously (e.g. fraudulent/erroneous entries)
  - Augment human decision-making with smart analytics to increase speed and accuracy (e.g. sentiment and topic identification for feedback)
- Increase job satisfaction
  - Reduce mundane tasks, increase engaging ones
  - Cultivate and deepen skillsets of the future

THE CURRENT STATE OF DIGITIZATION

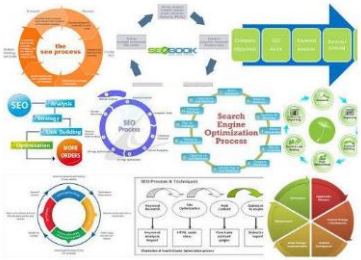




Source: McKinsey Global Institute "What's now and next in analytics, AI, and automation." May 2017

DIGITIZATION INVOLVES MORE THAN JUST TECH AND DEVELOPERS

- **New infrastructure and assets in technology**
  - Computing resources (machines, network etc.)
  - Data and the software to analyze them effectively

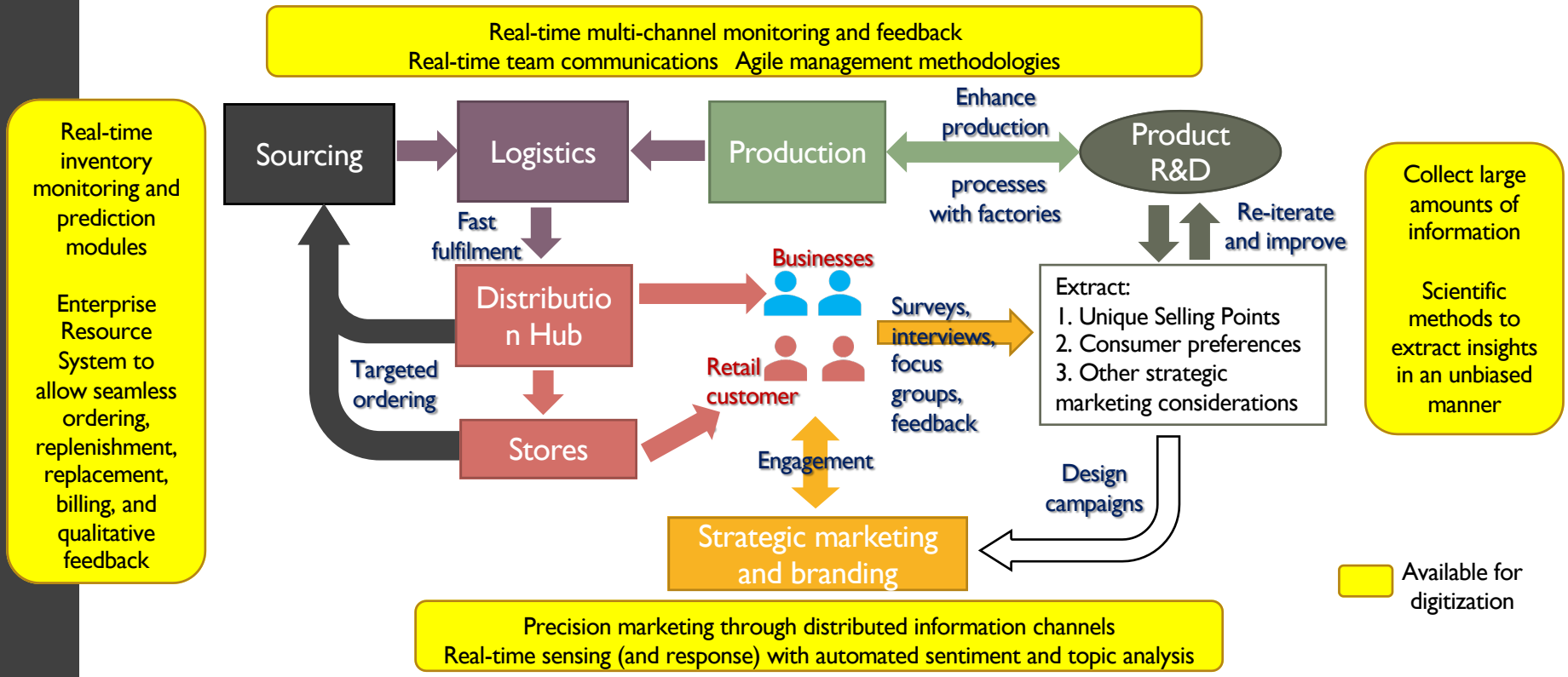


- **Re-engineer business management to new technology stack**
  - Workflow and processes
  - Operating, financing, investing activities
  - Upstream (supply chain) to downstream (customers) processes

- **Invest in human capital to wield new technology**
  - Enlightened management in technology
  - Key appointment holders that can bridge technology with business
  - Clearly defined digital roles and responsibilities

A REAL LIFE EXAMPLE (PRODUCT)

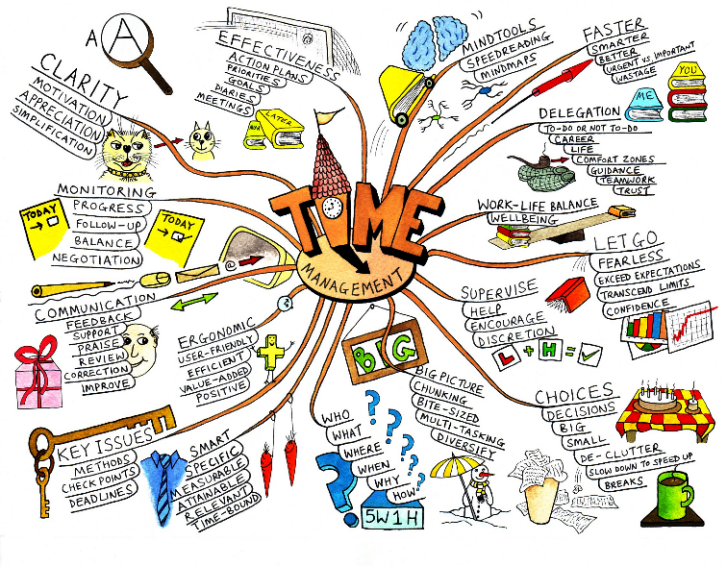
- Digitization involves hardware, software, business process re-design and most importantly, human capital



DEMYSTIFYING ANALYTICS

THE FOUNDATION OF ARTIFICIAL INTELLIGENCE

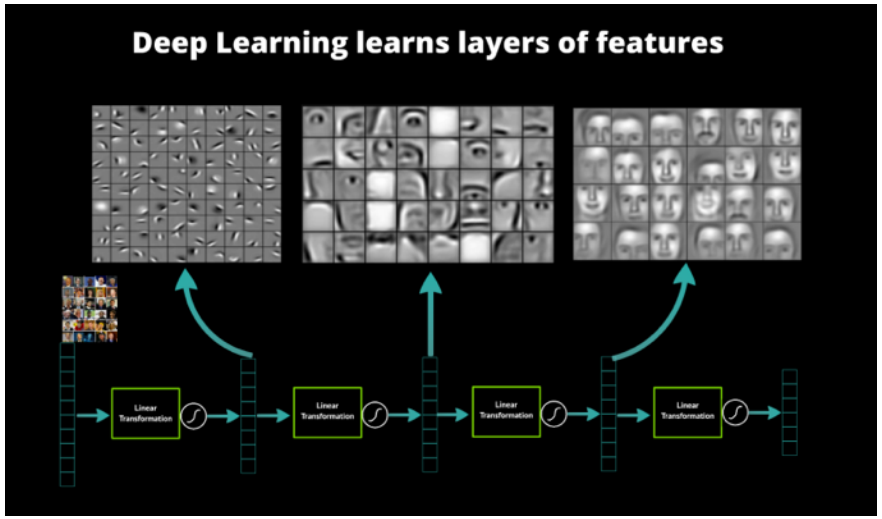
COMPUTERS VS HUMANS



- The common saying is that computers are way behind humans in higher cognitive skills, such as
  - Relationship mapping
  - Synthesizing different ideas
  - Extracting contexts

REALLY?

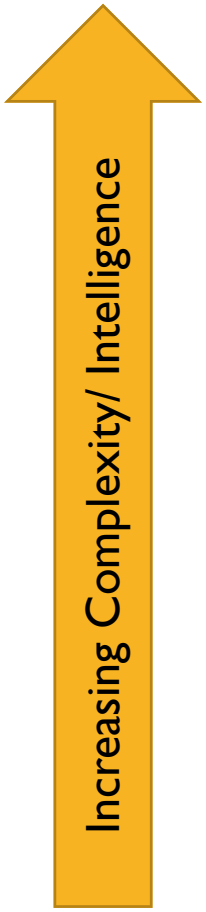
- Context-based AI
  - Siri and Google Assistant
- Bested the best experts in tasks that are considered to be too complex even for most humans
  - Defeated 7 times world champion in Jeopardy! (a game where contestants guess the question from 3 related answers)
  - IBM's Deep Blue defeated Garry Kasparov in 1997
  - Google's AlphaGo bested the best grandmasters in recent years in the game of Go
    - Google's AlphaGo Zero then beat AlphaGo by 100 games to 0



WE NEED TO BE EQUIPPED FOR A FUTURE BUILT BY DATA

- The core value-add of these artificial intelligence (AI) applications lies in the ability to make sense of what has happened, is happening, and predict what will happen
- The tools in which this core value-add is built upon
  - Insane amounts of data
  - A huge variety of data types (from quantitative to qualitative)
  - Statistics, machine learning, and other mathematical methods
- Computing resources and skillsets are critical because they are the only avenues for us to unlock the value of AI
- The future of data science is not tailored for only computer science specialists
  - Domain experts who can teach machines what and how to learn
  - Innovative people who knows how to use machines and AI to push the boundaries of the fields that they are in (including business)
  - **Analytics algorithms have been commoditized to the extent that anyone can use them without substantial mathematical or coding training**
  - **Analytics algorithms are general functions that are implemented in the same way across all analytics software (learn one, learn all)**

WHAT IS ANALYTICS?



- Relating to domain expertise
  - Asking the right questions (formulate hypotheses)
  - Extracting the right answers (interpreting results from analytical models)
- Relating to analytical techniques
  - Using machine learning and deep learning to extract patterns from data (without defining rules)
  - Using statistics to represent and describe the data
  - Applying decision rules to data
  - Aggregating and segmenting data (dice and slice)
- Relating to data technology
  - Translating business data into machine readable formats
  - Managing and linking data sources
  - Allowing large amounts of data to be stored and retrieved quickly

Analytics help us discover new insights from data,
through rigorous scientific interpretations of the patterns in them

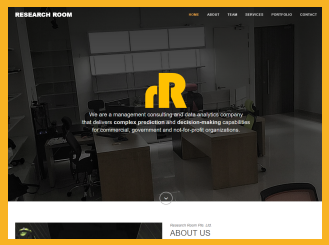
WHERE DO WE PLACE ANALYTICS IN THE DIGITIZATION VALUE CHAIN?

Frontend

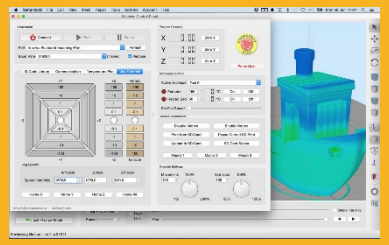
Focused on the development of user interfaces (with good User Experience)

UI: User Interface; UX: User Experience

Apps UI/UX



Web UI/UX



System UI (typically internal facing)

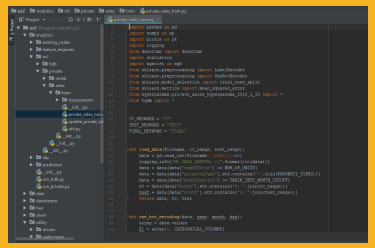
User Inputs

Results Output

Backend

Focused on the development of the digital architecture and functionalities

Architecture codes
- Glues all digital components together



Functionality codes
- Supports the functionalities of the digital solution

Analytics codes
- Creates intelligent responses based on **data inputs** from any components in the digital solution

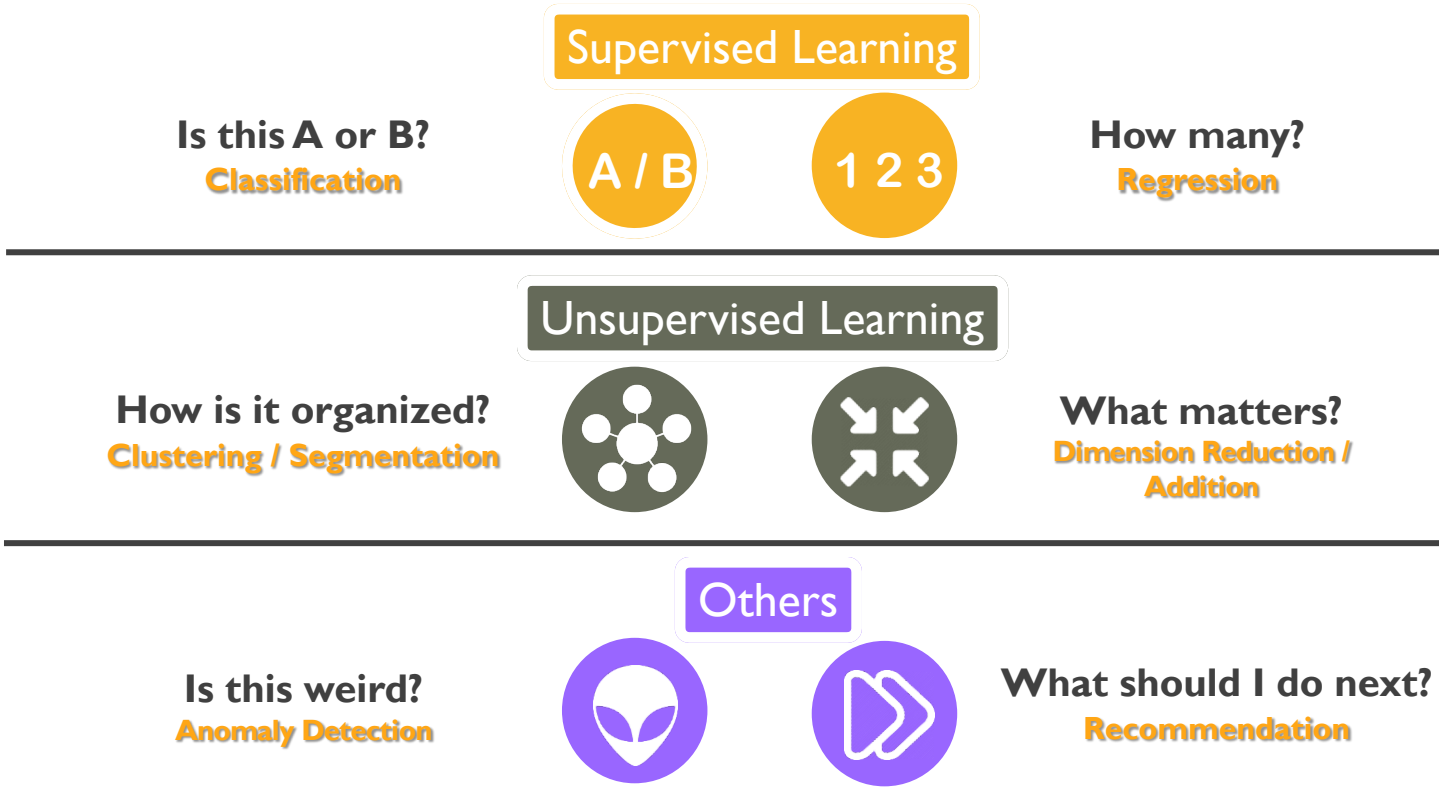
Database
Where all data are stored and persisted

In god we trust, all others bring data

- W. Edwards Deming

ANALYTICS ALGORITHMS: WHAT DO THEY DO?

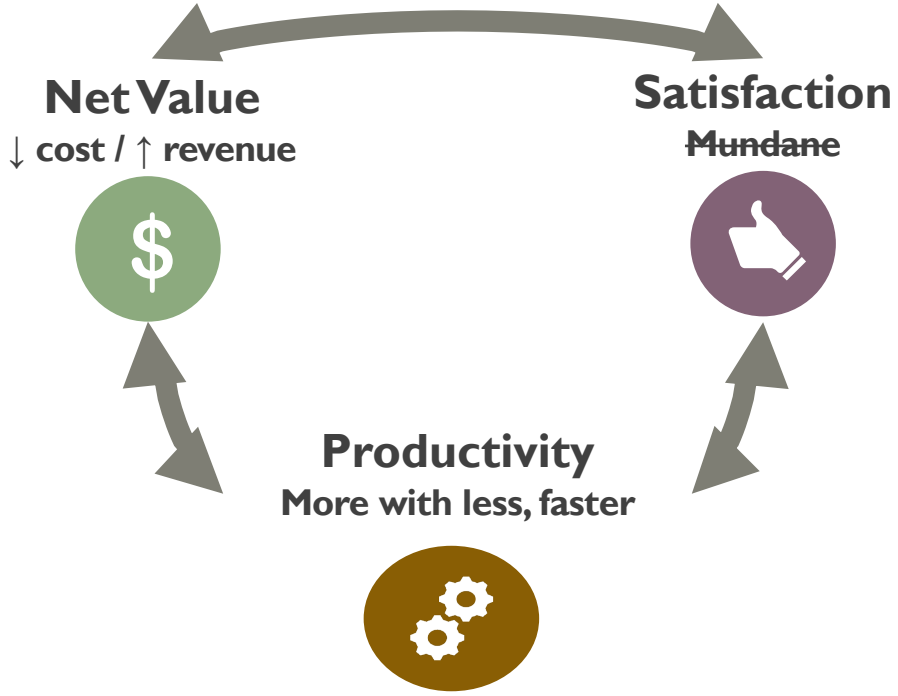
- Analytics algorithms uses mathematics to extract complex patterns from data
- These mathematical techniques are generally classified into:
  - Supervised learning: Humans show the machine which output relates to which inputs, and the machines will learn the associations without additional human intervention
  - Unsupervised learning: Identifies patterns in the data without any human intervention



ANALYTICS AND MUST BE GUIDED BY BUSINESS OBJECTIVES

- WHERE ARE YOU POINTING YOUR WEAPON AT?

- As a simple analogy
  - analytics algorithms tell you the best way to reach point A from point B
  - however, you need to determine what points A and B are
- Analytics is powerful when you have clear business objectives/outcomes
- Generally, analytics have shown great success in enhancing the following business objectives



ANALYTICS USE-CASES

NETFLIX RECOMMENDATION

- **Answering:**

- **Business Objective:**

- Netflix has classified 76,897 micro-genres for their collection of shows and movies
- Using browsing behaviors, Netflix is able to associate viewer characteristics with what they (or similar users) watched
  - A popular analytics algorithm that is used to for similar recommendation systems is the Matrix Factorization
- This probabilistic model is used to infer what new viewers may like to watch

COMMUNITY ENGAGEMENT

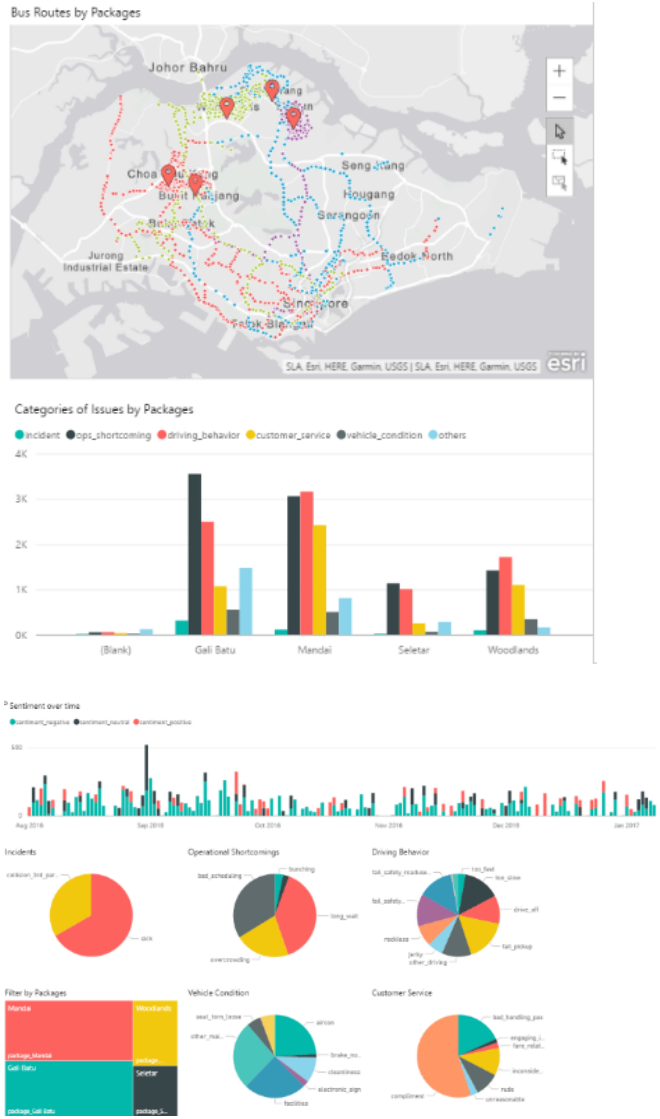
- **Answering:**

- **Business Objective:**

- This project utilizes the same recommendation algorithms on activities curation
- Depending on user profile and preferences, with environment variables (location, time of day, weather etc.)
  - The analytics algorithm can recommend immediate activities that the user may like to pursue

SOCIAL MEDIA ANALYTICS

- **Answering:** A / B classification
- **Business Objective:** productivity
- By associating words used in social media with sentiments and entities (e.g. air-conditioning, doors, seats)
  - The analytics algorithm is able to pick up negative posts, and extract important contexts within the post
  - Sentiment: complaint or compliment?
  - What is it about? (e.g. moving too slowly, air con too warm)
- Using analytics allow us to
  - Fully automate this customer service process
  - Provide consistent & high quality classification outcomes



PREDICTIVE TAXI DISPATCHING

- **Answering:** how many? 1 2 3
- **Business Objective:** productivity
- Machine learning algorithms was used to associate past taxi waiting times at Changi Airport Taxi Holding Area with environment variables such as
  - Weather
  - Time / Day (Holidays/Seasonal)
  - Flight load (no., origin)
- This model was then used to predict taxi waiting times
  - The basic model is accurate within 78.8 seconds, 95% of the time
- Unsupervised algorithms also revealed a network of taxi drivers who can consistently achieve very little waiting time
  - The existence of taxi driver cliques with a common communication network

TARGET CORP MARKET BASKET ANALYTICS

- **Answering:**

segmentation

classification

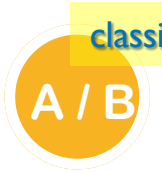
- **Business Objective:**

net value

- This is a famous example of the market basket algorithm in machine learning
- Target used the algorithm on transaction level data and computed the probabilities of 2 or more items appearing in the same customer checkout cart
  - If the actual probability of 2 items appearing in the same cart is higher than the joint probability, then this pair of products is considered novel
- Target then places novel products in close proximity within their stores to drive sales
  - Beer & Diapers
  - Scent-free soap, extra-big bags of cotton balls

URBAN ZOOM – HOME AUTO VALUATION

- **Answering:**



classification

segmentation

- **Business Objective:**

satisfaction

- Using advanced machine learning models on a massive data set that has every unit-level transaction in Singapore since 1980
  - associate home prices with a huge database of home, location, and time characteristics
  - Able to facilitate instant valuation for all homes in Singapore
- The model is able to achieve median error in predicted value of 2.4%

CUSTOMIZED EDUCATION JOURNEY

- **Answering:**

segmentation

classification

- **Business Objective:**

satisfaction

- Education content is mapped to a knowledge map based on a financial curriculum
- Advanced machine learning algorithms predict the optimal learning journey and recommend new content based on user interests (as revealed by their choice of content)

Stock-Picking: Skill or Luck?

INVEST 3 min read

By Dr. Jack Hong
08 Mar 2018

COPY LINK

Before we explore the logic in the title, let's start with a simple experiment: a game of coin toss.

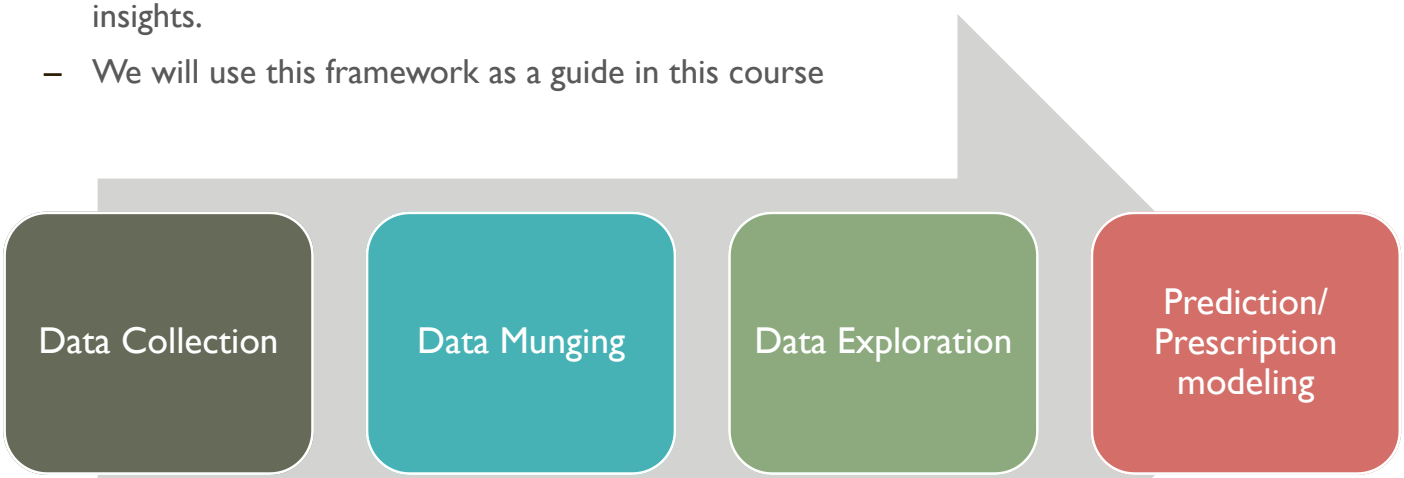
Let's assume you start with a wealth of $100 and perform a sequence of coin tosses. When the coin lands on heads, you get a 3% increment to your wealth, and when it lands on tails, you lose 3% of your

DATA SCIENCE FRAMEWORK

FROM IDEAS TO DELIVERABLES

THE DATA SCIENCE VALUE CHAIN

- The process to bring analytics ideas to life can be generalized into an empirical framework
  - This framework is modeled after the scientific approach that empirical researchers use to develop insights.
  - We will use this framework as a guide in this course



- Identify data sources
- Collect data (APIs, crawlers, human labor etc.)
- Identify and implement the best data structures for storage and retrieval

- Identify and access datasets
- Append/merge datasets into views
- Clean, transform and replace data
- Descriptive statistics and preliminary data investigation

- Apply multiple statistical and machine learning models
- Model selection
- Model comparison

- Analyze useful features
- Does it fit the research questions?
- Predict the future using the best model
- Consider budget and other constraints

HEAVILY RELIANT ON COMPUTING POWER

- Each element in the data science value chain is heavily reliant on computing power
- Step 1: Data collection
  - Many information sources are in data format (e.g. excel, csv)
  - Increasingly, data are being provided in API format (e.g. data.gov.sg) which requires coding for retrieval purposes
  - Some data sources are so massive that we could only pull in batches or on demand
- Step 2: Data munging
  - Many data sources are not represented in a form that fits our purposes
  - Many data sources are also "dirty", such as data entry errors, missing data, wrong formatting
  - Due to the sheer size of the data that we need to work with, manual cleaning takes too much time
  - As such, we rely on codes and algorithms to clean the data
- Step 3 and 4: Data exploration, prediction, and prescription
  - Computing by hand is not an option

COMPUTING BASICS

WHAT IS A COMPUTER DO?

COMPUTING BASICS (HARDWARE)

- What does a computer do?
  - A computer computes!

- Using the biological brain as an analogy

Central Processing Unit (CPU)
=
Frontal Lobe (calculates)

GPU (graphical processing unit) is similar to CPU and is extensively used in deep learning models

1 video card packs hundreds to thousands of processing units, compared to 2 or 4 processing units in a laptop/personal computer

Random Access Memory (RAM)
=
Working (fast) memory

Accessing data from RAM is much faster than accessing data from disks

However, all data in RAM is flushed out when the machine is turned off

Also, RAM is much more expensive than disks

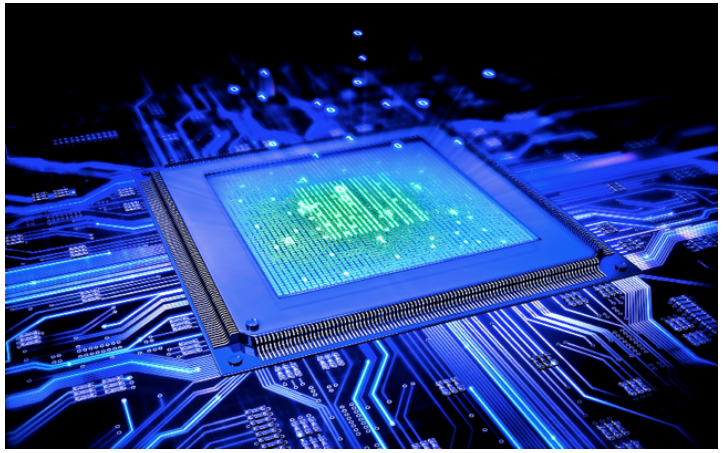
Disk Drives
=
Long term (slow) memory

This is the cheapest form of data storage

Data persists in these disks until you scrub them

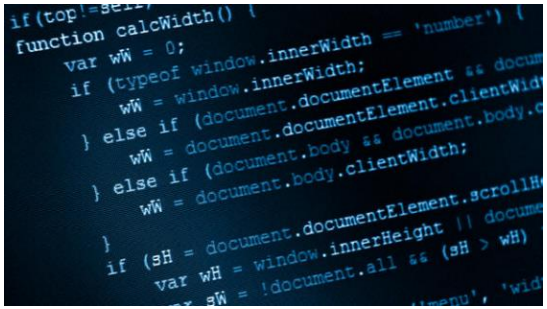
COMPUTERS VS HUMANS

- Computers beat brains in terms of
  - Speed (computation and recall)
    - How long do you think it will take my MacBook Pro to compute (369123 * 369123) and repeat it **100 million (100,000,000)** times?
    - Answer: Less than 4 seconds
  - Accuracy (perfect recall)
    - Will a computer system return different data the same way you call it every time?
  - Working volume
    - How many wikipedia (the entire site with all its data) can a S$100 hard disk hold?
    - Answer: About 10



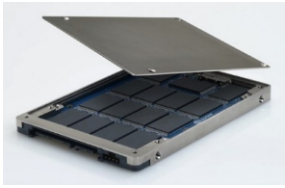
ARCHITECTURE

- What happens when you execute a piece of code?
  - CPU computes according to programming logic
    - Operators: + - / x
    - Logic: if/then/else
    - Flow: while x < 1



  - Where does it store and read data and steps?
    - Random Access Memory (RAM)
    - Requires unique physical locations on the hardware (memory addresses)

  - Since RAM is like a storage, can we use disk drives in place of RAM?
    - Yes, OS X and Windows are doing that (virtual memory/pagefile/swap)
    - But they only use it for very specific situations, why?
      - RAM is 33x faster than SSD, which is 100x faster than HDD
      - Disk drives have limited writes (SSD will wear out in 2 months)
      - CPUs can only pre-load data into its cache (small temporary storage area) from RAM



QUESTIONS?

Email any queries to
jackhong@smu.edu.sg